



Wilkie, C., Miller, C., Scott, M., Simis, S., Groom, S., Hunter, P., Spyarakos, E. and Tyler, A. (2018) Spatiotemporal Statistical Downscaling for the Fusion of In-lake and Remote Sensing Data. In: 33rd International Workshop on Statistical Modelling (IWSM 2018), Bristol, UK, 16-20 Jul 2018, pp. 207-212.

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

<http://eprints.gla.ac.uk/169725/>

Deposited on: 27 September 2018

Enlighten – Research publications by members of the University of Glasgow\_  
<http://eprints.gla.ac.uk>

# Spatiotemporal statistical downscaling for the fusion of in-lake and remote sensing data

Craig Wilkie<sup>1</sup>, Claire Miller<sup>1</sup>, Marian Scott<sup>1</sup>, Stefan Simis<sup>2</sup>,  
Steve Groom<sup>2</sup>, Peter Hunter<sup>3</sup>, Evangelos Spyarakos<sup>3</sup>, Andrew  
Tyler<sup>3</sup>

<sup>1</sup> University of Glasgow, UK

<sup>2</sup> Plymouth Marine Laboratory, UK

<sup>3</sup> University of Stirling, UK

E-mail for correspondence: [craig.wilkie@glasgow.ac.uk](mailto:craig.wilkie@glasgow.ac.uk)

**Abstract:** This paper addresses the problem of fusing data from in-lake monitoring programmes with remote sensing data, through statistical downscaling. A Bayesian hierarchical model is developed, in order to fuse the in-lake and remote sensing data using spatially-varying coefficients. The model is applied to an example dataset of log(chlorophyll-*a*) data for Lake Erie, one of the Great Lakes of North America.

**Keywords:** Bayesian hierarchical model; Statistical downscaling; Data fusion; Chlorophyll-*a*.

## 1 Introduction and background

This work is motivated by the problem of fusing data from in-lake monitoring programmes with remote sensing data, which have impressive spatial and temporal coverage but require calibration with the in-lake data to ensure accuracy. This presents a problem of change-of-support between the point-scale in-lake data and the grid-cell-scale remote sensing data.

In-lake data have been traditionally used extensively to enable water quality investigators to understand lake health. They are assumed to be accurate within measurement error, since they are obtained from water samples that are taken directly from the lake surface and then analysed in a laboratory. However, these data are expensive to collect in terms of both time and money and so are often sparse in both space and time, with a small

---

This paper was published as a part of the proceedings of the 33rd International Workshop on Statistical Modelling (IWSM), University of Bristol, UK, 16-20 July 2018. The copyright remains with the author(s). Permission to reproduce or extract any parts of this abstract should be requested from the author(s).

number of sampling locations across each lake. They therefore provide little information on the spatial patterns in water quality. Remote sensing data have become much more commonly available in recent years, due to the increased availability of data from Earth-facing satellite monitoring programmes. These data provide spatially comprehensive information on water quality parameters.

In this paper, data for  $\log(\text{chlorophyll-}a)$ , an important indicator of lake water quality, are considered. The example used is Lake Erie, one of the Great Lakes of North America, which has suffered from poor water quality in the past and is therefore of interest to regulatory bodies and local communities. The in-lake data are available for 20 locations over 20 months, collected by the US Environmental Protection Agency and made available in the LIMNADES database (<https://www.limnades.org/home.psp>). These data are collected at several time points within each month and are temporally aggregated onto the monthly scale, before analysis. The remotely-sensed data are available over the same time period, but with a much better spatial coverage, with grid cells of up to 300 metres in dimension, with 351,041 grid cells covering the lake, on a monthly-averaged time-scale. These European Space Agency Medium Resolution Imaging Spectrometer data were produced through the GloboLakes project and are available at <https://globolakes.eofrom.space/>.

The remotely-sensed data and the in-lake data for August 2007 are shown in Figure 1 below.

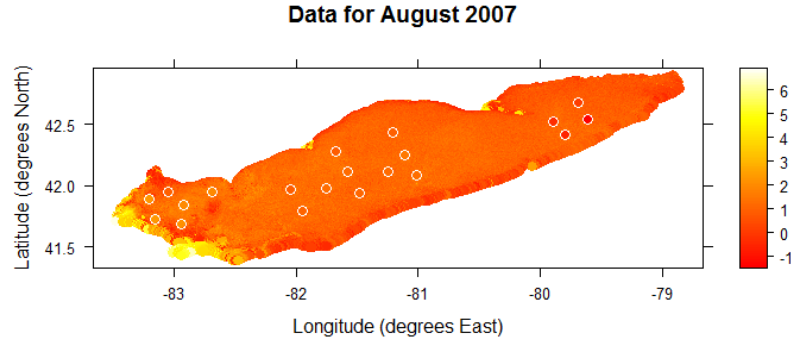


FIGURE 1. Remote sensing data for August 2007, with the in-lake data overlaid and surrounded by white circles.

This paper presents a spatiotemporal development of the model of Wilkie et al. (2015), with an application to a spatially-larger dataset. The model is based upon the approach of Gelfand et al. (2003), which was developed into a statistical downscaling model by Berrocal et al. (2010) for air quality data.

## 2 Methodology

A Bayesian hierarchical model is proposed for the fusion of remote sensing and in-lake data. The model allows for the  $n_j$  in-lake sampling locations to differ for each time point  $j$  (for  $j = 1, \dots, t$ ). For the vector of response data  $\mathbf{y}_j$ , i.e. the vector of in-lake data collected at the  $n_j$  sampling locations at time  $j$ , and the vector of remote sensing data  $\mathbf{x}_j$  recorded for the  $n_j$  grid cells containing these in-lake sampling locations, the model is written as follows:

$$\mathbf{y}_j \sim N_{n_j}(\boldsymbol{\alpha}_j + \boldsymbol{\beta}_j \odot \mathbf{x}_j, \sigma_\epsilon^2 \mathbf{I}_{n_j}),$$

where the vectors of intercepts and slope coefficients  $\boldsymbol{\alpha}_j$  and  $\boldsymbol{\beta}_j$  are allowed to be smoothly spatially-varying and so are given the following multivariate-Normal prior distributions:

$$\boldsymbol{\alpha}_j \sim N_{n_j}(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \mathbf{D}_j)) \text{ and } \boldsymbol{\beta}_j \sim N_{n_j}(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \mathbf{D}_j)),$$

where  $\sigma_\alpha^2$  and  $\sigma_\beta^2$  are the spatial variance parameters and  $\phi_\alpha$  and  $\phi_\beta$  are the spatial decay parameters, controlling how fast the correlations in the intercept and slope parameters decrease to zero as the distance between the in-lake sampling locations increases. These parameters are shared over time, which helps to improve their estimation. The matrix  $\mathbf{D}_j$  is the  $n_j \times n_j$  matrix of distances between the in-lake sampling locations for time  $j$ . Finally, the remaining prior and hyperprior distributions must be specified. The spatial variance parameters and error variance parameter are given the following distributions:

$$\sigma_\alpha^2 \sim \text{Inv-Gamma}(2, 1), \sigma_\beta^2 \sim \text{Inv-Gamma}(2, 1) \text{ and } \sigma_\epsilon^2 \sim \text{Inv-Gamma}(2, 1),$$

following the example of Sahu et al. (2006). As noted by Sahu et al. (2006), the spatial decay parameters are not easy to identify and so a grid search is performed.

All full conditional posterior distributions can be derived. Therefore, the model is fitted using Gibbs sampling. To provide predictions over the lake surface, Delaunay triangulation, constrained by the lake edges, is carried out to ensure the optimal spatial coverage of the prediction locations.

The temporal aspect of the data is made use of through the sharing of information over time, with the error variance and spatial variance parameters being estimated from the data for all timepoints.

## 3 Example for Lake Erie

Using the example dataset for Lake Erie, the model is fitted using the R packages `Rcpp` and `RcppArmadillo`, with predictions made at 1000 locations for each of the 20 months in the dataset. These locations are defined by a Delaunay triangulation that is constrained by points along the lake

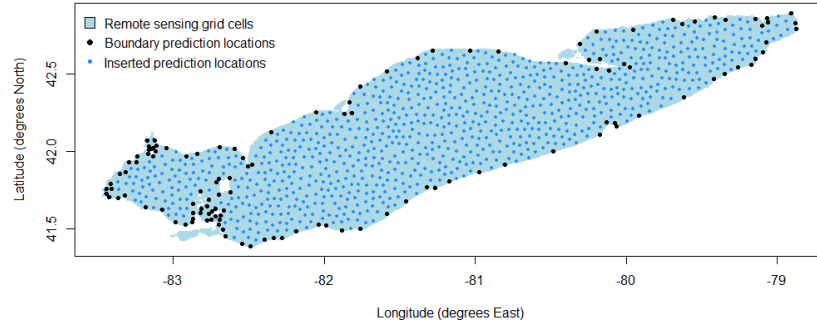


FIGURE 2. Remote sensing grid cells with prediction points overlaid, as obtained from a Delaunay triangulation constrained by the boundary points.

edges, using the R package `RTriangle`. The constraining points and the resulting inserted points are shown in Figure 2.

The model is run for 2 chains of 10,000 iterations each, with every tenth iteration saved, after a burn-in period of 100 iterations. Trace and density plots, such as the examples for the prediction at prediction location 1 for August 2007 ( $\tilde{y}_{1,\text{Aug } 2007}$ ) shown in Figure 3, provide no evidence against the assumption that the MCMC chains have converged to their posterior distributions.

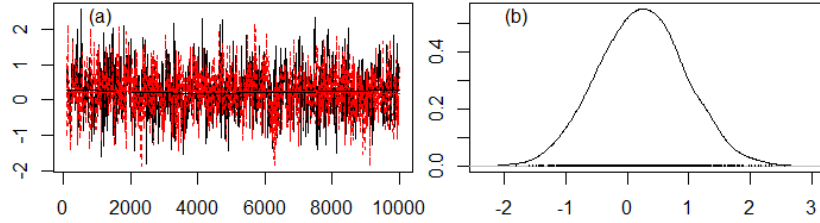


FIGURE 3. (a) Trace plot of the MCMC iterations for  $\tilde{y}_{1,\text{Aug } 2007}$ ; (b) Plot of the posterior density for  $\tilde{y}_{1,\text{Aug } 2007}$ .

The resulting predictions for August 2007 are shown in Figure 4(a) and their corresponding standard errors are given in Figure 4(b). These predictions illustrate the utility of the model for calibrating the remotely-sensed data using the in-lake data, while retaining the important spatial patterns of the remote sensing data. Figure 4(a) shows the adjustments to the remote sensing image of Figure 1 as a result of the fusion with the in-lake data. In the example shown here, the model predictions show that the northeast of the lake has lower values of  $\log(\text{chlorophyll-}a)$  in August 2007, while the southwest of the lake has higher values. Figure 4(b) shows that the standard errors are lowest closest to the in-lake data locations for this

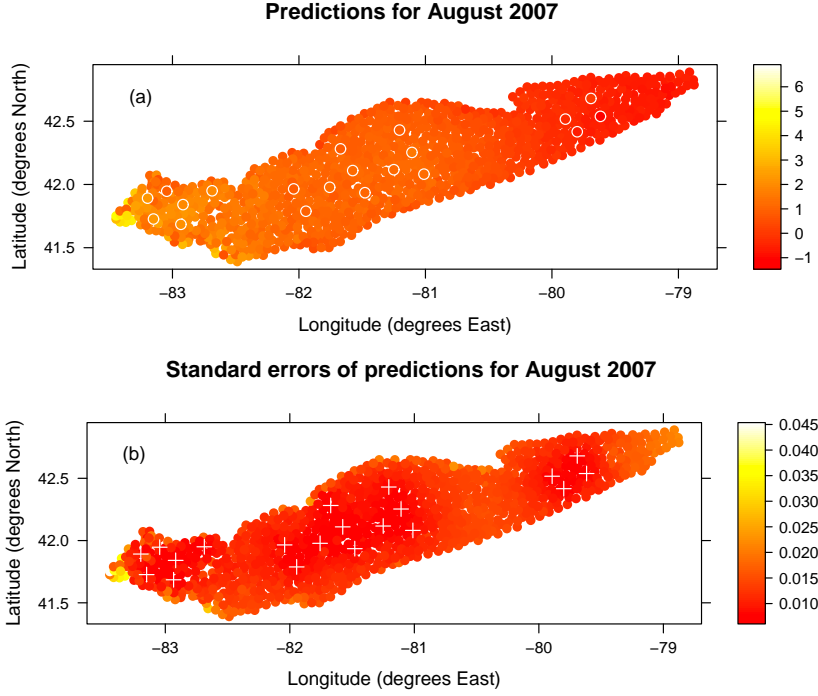


FIGURE 4. (a) Predictions for August 2007, with the in-lake data overlaid and surrounded by white circles; (b) Standard errors of predictions for August 2007, with the in-lake data locations marked by white crosses.

month, as expected. The standard errors are small in comparison to the variation across the lake, providing evidence of a true pattern across the lake surface. The resulting spatial maps, such as the example shown in Figure 4(a), would be useful for water quality investigators to identify parts of the lake of particular interest for further study.

## 4 Conclusions

The model described in this work enables the fusion of data from in-lake monitoring schemes, which are limited spatially and temporally, with extensive remotely-sensed data with good spatial and temporal resolution. The model makes use of data from multiple available timepoints in order to improve the estimation of the spatial variance parameters and the error variance parameter. Predictions can be made at any point location for which corresponding remotely-sensed data are available, i.e. any location within a remote sensing grid cell. Delaunay triangulation is used to optimise the spatial coverage of the prediction locations, in order to gain a better

understanding of the state of the health of the lake without increasing the computational complexity of the model.

Future work focusses on dealing with the temporal change of support, which can be accomplished through treating the data for each in-lake location and remote-sensing grid-cell as observations of smooth functions over time.

**Acknowledgments:** CW acknowledges the support of the School of Mathematics and Statistics, University of Glasgow, for funding the PhD during which this work was carried out. CM and MS were partly funded for this work through the NERC GloboLakes project (NE/J022810/1).

## References

- Berrocal, V.J., Gelfand, A.E., and Holland, D.M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological and Environmental Statistics*, **15**, 176–197.
- Gelfand, A.E., Kim, H.-J., Sirmans, C.F., and Banerjee, S. (2003). Spatial modelling with spatially-varying coefficient processes. *Journal of the American Statistical Association*, **98**, 387–396.
- Sahu, S.K., Gelfand, A.E., and Holland, D.M. (2006). Spatio-temporal modelling of fine particulate matter. *Journal of Agricultural, Biological and Environmental Statistics*, **11**, 61–86.
- Wilkie, C.J., Scott, E.M., Miller, C., Tyler, A.N., Hunter, P.D., and Spyarakos, E. (2015). Data fusion of remote-sensing and in-lake chlorophyll<sub>a</sub> data using statistical downscaling. *Procedia Environmental Sciences*, **26**, 123–126.